BMC Proceedings

PROCEEDINGS Open Access

A penalized linear mixed model for genomic prediction using pedigree structures

Can Yang[1], Cong Li[2], Mengjie Chen[2], Xiaowei Chen[2], Lin Hou[1], Hongyu Zhao[1,2*]

From Genetic Analysis Workshop 18
Stevenson, WA, USA. 13-17 October 2012

Abstract

Genetic Analysis Workshop 18 provided a platform for evaluating genomic prediction power based on single-nucleotide polymorphisms from single-nucleotide polymorphism array data and sequencing data. Also, Genetic Analysis Workshop 18 provided a diverse pedigree structure to be explored in prediction. In this study, we attempted to combine pedigree information with single-nucleotide polymorphism data to predict systolic blood pressure. Our results suggested that the prediction power based on pedigree information only could be unsatisfactory. Using additional information such as single-nucleotide polymorphism genotypes would improve prediction accuracy. In particular, the improvement can be significant when there exist a few single-nucleotide polymorphisms with relatively larger effect sizes. We also compared the prediction performance based on genome-wide association study data (ie, common variants) and sequencing data (ie, common variants plus low-frequency variants). The experimental result showed that inclusion of low frequency variants could not lead to improvement of prediction accuracy.

Background

Genomic prediction is an important problem in genetics. It aims at predicting a phenotype outcome based on information from genetic markers, population, pedigree structures, and other relevant covariates. Recent studies suggest that genomic prediction based on genome-wide case control data (unrelated individuals) has limited prediction accuracy [1]. First, the difficulty may be caused by the polygenicity of complex traits, that is, many markers with small effects jointly affect the trait [2-4]. A larger sample size is needed to estimate those small effects more accurately. A larger sample size also leads to the improvement of prediction accuracy. Second, low frequency variants (minor allele frequency [MAF] ≤5%) have not been directly observed in genome-wide association studies (GWAS). The contribution of these low-frequency variants has not been taken into account in predictive models, which may result in the loss of prediction accuracy.

Genetic Analysis Workshop 18 (GAW18) provides both genotyping data and sequencing data of approximately 1000 samples from 20 pedigrees. This brings a good opportunity to evaluate genomic prediction from the following 2 perspectives:

1. How do we integrate the pedigree structures and genome-wide dense markers for evaluating the power of genomic prediction?
2. Can we improve prediction accuracy by including low-frequency variants?

Some pioneering studies suggested that integration of phenotype information from relatives and informative markers can improve prediction accuracy [5-7]. Linear mixed models (LMMs) have arisen as a useful tool for information integration in this context [8,9]. The random effects can used to model pedigree and the fixed effects can be used to include informative markers. It is difficult to use LMM when the number of fixed effects exceeds the number of samples. To overcome this difficulty, bayesian linear regression [6] and bayesian alphabet methods [7] have been proposed. Alternatively, an L_1 estimation procedure

* Correspondence: hongyu.zhao@yale.edu
[1]Department of Biostatistics, Yale School of Public Health, New Haven, CT 06520, USA
Full list of author information is available at the end of the article

has been proposed for LMM, named "LMMLasso" [8]. Very recently, this model has been applied to association mapping, where the random effects were used for population stratification correction [9]. However, it is computationally too intensive to apply either LMMLasso or bayesian linear regression to the genome-wide scale data set from GAW18. The aim of this study is to provide an efficient computational method to evaluate genomic prediction and answer the 2 questions above.

Methods

Basic model

Let n be the sample size. We consider the following LMM

$$\begin{aligned} \mathrm{y} &= \mathrm{X}\beta + \mathrm{G}\alpha + u + e, \\ \mathrm{u} &\sim N(0, \sigma_u^2 \mathrm{K}), \\ \mathrm{e} &\sim N(0, \sigma_e^2 \mathrm{I}), \end{aligned} \tag{1}$$

where $\mathbf{y} \in R^{n\times 1}$ is the response vector; $\mathbf{X} \in R^{n\times d}$ is the matrix of covariates (fixed effects), including the intercept and other covariates, such as age and gender; $\boldsymbol{\beta}$ is the vector for regression coefficients of the covariates; $\mathbf{G} \in R^{n\times p}$ is the genotype matrix and $\boldsymbol{\alpha}$ is the coefficient vector for p single-nucleotide polymorphisms (SNPs) (fixed effects); $\mathbf{u}$ is the random effect from $N(0,\sigma_u^2\mathbf{K})$; and $\mathbf{e}$ is the residual error with variance σ_e^2. Here the covariance matrix $\mathbf{K}$ is the genetic relatedness matrix that describes the pedigree structure among the individuals. The covariance matrix $\mathbf{K}$ can be constructed according to the known pedigree information or estimated from genome-wide SNP information.

Penalized linear mixed model

There is a difficulty in applying the model when d+p+2>n,ie, the number of parameters exceeds the number of samples (d is the number of covariates, p is the number of SNPs treated as fixed effects, 2 is the number of variance components). To overcome this difficulty, we use penalized LMM to perform model selection [8,9]. Consider introducing a penalty on the coefficient $\boldsymbol{\alpha}$: $P(\boldsymbol{\alpha}) \leq t$, where $P(\boldsymbol{\alpha})$ is the penalty and t is some constant. First, we can write down the log-likelihood of LMM (1) by integrating out $\mathbf{u}$ and $\mathbf{e}$ as

$$\begin{aligned} L(\sigma_e^2, \sigma_u^2, \beta, \alpha) &= \log N(\mathbf{y}|\mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\alpha}; \sigma_u^2\mathbf{K} + \sigma_e^2\mathbf{I}) \\ &= \log N(\mathbf{y}|\mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\alpha}; \sigma_u^2(\mathbf{K} + \delta\mathbf{I})) \end{aligned} \tag{2}$$

where $\delta = \sigma_e^2/\sigma_u^2$. By eigendecomposition, $\mathbf{K} = \mathbf{USU}^T$. After some algebraic operations, we have

$$L(\delta, \sigma_u^2, \boldsymbol{\beta}, \boldsymbol{\alpha}) = -\frac{1}{2}\left(n\log(2\pi\sigma_u^2) + \log\det(\mathbf{S} + \delta\mathbf{I}) + \frac{1}{\sigma_u^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{G}\boldsymbol{\alpha})^T\mathrm{U}(\mathbf{S} + \delta\mathbf{I})^{-1}\mathrm{U}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{G}\boldsymbol{\alpha})\right)$$

To maximize the above log-likelihood with the constraint $P(\boldsymbol{\alpha}) \leq t$, equivalently, we may minimize the Lagrange form of the negative log-likelihood:

$$\min_{\delta,\sigma_u^2,\boldsymbol{\beta},\boldsymbol{\alpha}} \frac{1}{2}\left(n\log(2\pi\sigma_u^2) + \log\det(\mathbf{S} + \delta\mathbf{I}) + \frac{1}{\sigma_u^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{G}\boldsymbol{\alpha})^T\mathbf{U}(\mathbf{S} + \delta\mathbf{I})^{-1}\mathbf{U}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{G}\boldsymbol{\alpha})\right) + \lambda P(\boldsymbol{\alpha}) \tag{3}$$

To optimize the penalized log-likelihood function, we adopt an alternating strategy as follows:

Step 1: For fixed $\boldsymbol{\alpha}$, we can treat $\mathbf{y} - \mathbf{G}\boldsymbol{\alpha}$ as the working response and use maximum likelihood (ML) or restricted maximum likelihood (REML) to obtain ($\boldsymbol{\beta}$, σ_u^2, δ) using some recent algorithms proposed to efficiently solve the optimization, such as FastLMM [10] and GEMMA [11].

Step 2: For fixed ($\boldsymbol{\beta}$, σ_u^2, δ), the problem (3) becomes

$$\begin{aligned} &\min_{\boldsymbol{\alpha}} \left(\frac{1}{2\sigma_u^2}(\mathbf{U}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{G}\boldsymbol{\alpha}))^T(\mathbf{S} + \delta\mathbf{I})^{-1}\left(\mathbf{U}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{G}\boldsymbol{\alpha})\right)\right) + \lambda P(\boldsymbol{\alpha}) \\ &= \min_{\boldsymbol{\alpha}} \frac{1}{2\sigma_u^2}\left\|\tilde{\mathbf{y}} - \tilde{\mathbf{G}}\boldsymbol{\alpha}\right\|^2 + \lambda P(\boldsymbol{\alpha}) \end{aligned} \tag{4}$$

where $\tilde{\mathbf{y}} = (\mathbf{S} + \delta\mathbf{I})^{-\frac{1}{2}}\mathbf{U}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ and $\tilde{\mathbf{G}} = (\mathbf{S} + \delta\mathbf{I})^{-\frac{1}{2}}\mathbf{U}^T\mathbf{G}$. When we choose $P(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_1$, the optimization problem (4) becomes the standard Lasso problem [12]. In fact, we may have some other choices of penalties, such as the elastic net [13] and MC+ penalties [14]. Coordinate descent algorithms can be used to solve the penalized regression problem efficiently [14,15].

Implementation details

Because the optimization problem (3) is not convex, a good initial point will help to find a better solution. We started at $\boldsymbol{\alpha} = \mathbf{0}$, and used REML to initialize ($\boldsymbol{\beta}$, σ_u^2, δ). As in Friedman et al [15], we used $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{G}}$ to calculate the smallest λ value such that all $\boldsymbol{\alpha}$ equal to zero. We denote this λ as λ_0. Then we generated a decreasing sequence of λ values such that $\lambda_{i+1} = \eta\lambda_i, i = 0, 1, ...k$, where k is the length of the λ sequence. Let $(\boldsymbol{\alpha}, \beta, \sigma_u^2, \delta)_i$ be the solution corresponding to λ_i. When we were solving the $(\boldsymbol{\alpha}, \beta, \sigma_u^2, \delta)_{i+1}$ for the regularization parameter λ_{i+1}, we always used $(\boldsymbol{\alpha}, \beta, \sigma_u^2, \delta)_i$ as the initial point. This strategy accelerates the convergence of the algorithm. Typically, we set $\eta = 0.95$, $k = 20$, and choose the best λ value by cross-validation.

Results

Data set and preprocessing

For the GAW18 data set, there are 959 samples from 20 pedigrees. In our study, we considered systolic blood pressure (SBP) as the quantitative trait of interest. Among all the samples, 849 individuals had at least one blood pressure measurement. We considered the first nonmissing measurement of SBP and used its log-transformed value as the response $\mathbf{y} \in R^{849}$, with the age at the corresponding measurement as the covariate. We included the intercept as a fixed effect, giving us $\mathbf{X} \in R^{849\times 2}$.

For the genotype data matrix $\mathbf{G}$, we focused on genetic markers on chromosome 3. The reason we chose chromosome 3 was that we identified a significant signal in the gene *MAP4* by association tests on all 200 simulated SBPs. This signal was significantly stronger than the genetic background signal. We expected our method could automatically detect it and include it as a fixed effect.

We used both GWAS data and sequencing data to evaluate the model performance. For GWAS data, we applied basic quality control (MAF >0.05, missing rate <0.05), which resulted in 33,248 SNPs for chromosome 3 (ie, $\mathbf{G}_{gwas} \in R^{849\times 33248}$). For sequencing data, there were 602,512 SNPs for chromosome 3. We first applied the same quality control criteria and then did linkage disequilibrium pruning using the threshold r^2 = 0.9, leaving 103,020 SNPs for chromosome 3 (ie, $\mathbf{G}_{seq} \in R^{849\times 103020}$). The matrix $\mathbf{K}$ is twice the kinship matrix which is constructed based on pedigree information.

Results

Before we applied the proposed method for prediction, we first estimated the variance that can be explained by the pedigree structure. We used LMM to do this, and included age and intercept as the fixed effects and two times kinship matrix as the random component [2]. The estimate of explained variance was 26.98% and 18.85%, for the simulated and real phenotypes, respectively. This could be an overestimate because members within the same pedigree might share some environmental factors.

We applied penalized LMM to both GWAS data and sequencing data to evaluate its performance on phenotype prediction. Here we chose the Lasso penalty (ie, $P(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_1$). The results are shown in Figures 1 and 2. We ran 10-fold cross-validation 20 times to generate these boxplots. We reported the result for the λ sequence, λ_i, $i = 0, 1, \ldots 20$. When $\lambda = \lambda_0$, the model corresponds to LMM without genetic markers. As λ_i

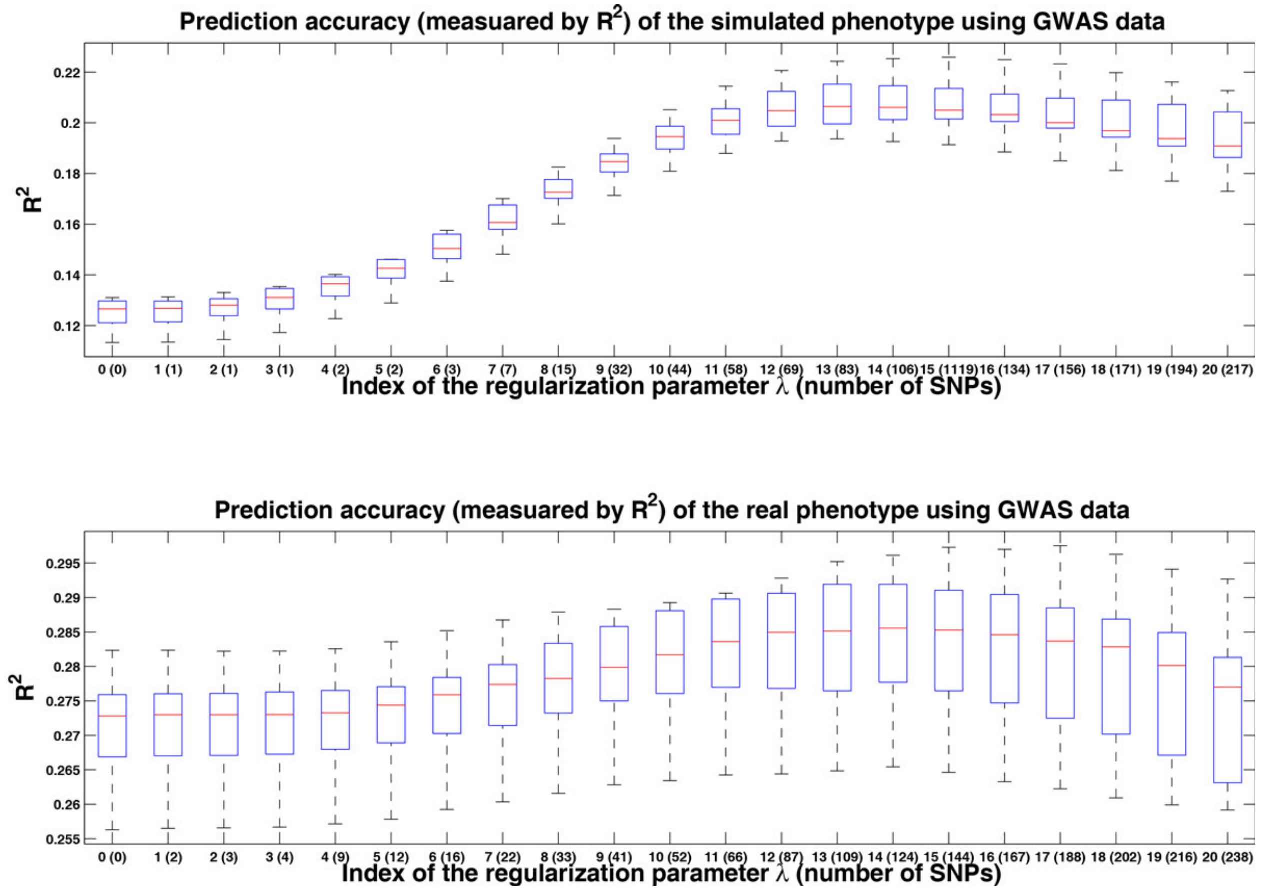

Figure 1 Prediction accuracy using GWAS data. Prediction results for the simulated phenotype and real phenotype using GWAS data. We ran 10-fold cross-validation 20 times to generate these boxplots. The x-axis is the index of the regularization parameter λ. The corresponding number of SNP markers used in the prediction model is given in brackets. Notice that these numbers are obtained using all samples. Index 0 corresponds to prediction only using pedigree information, thus the corresponding number of SNP marker is zero. The y-axis is the R^2 between the true value and the prediction value.

decreased, more and more genetic markers were used in the model.

Based on the estimated variance components (26.98% for the simulated SBP and 18.85% for the real SBP), it seemed that we should do a better job for the simulated phenotype. From Figures 1 and 2, however, we observed that we had a better performance for the real phenotypes (R^2 = 0.205 for the simulated SBP and R^2 = 0.285 for the real SBP). The reason was that the covariate "age" contributed more for the real phenotypes. If we did prediction for both the simulated and real phenotypes without "age," we could observe a slightly better performance for the simulated phenotype.

Let us take a close look at the result using GWAS data. We used the correlation between the true value and predicted value to measure the accuracy. For the simulated phenotypes, the accuracy is around R^2 = 0.125 for $\lambda = \lambda_0$, and kept improving until $\lambda = \lambda_{14}$. After that, the performance started to get worse. We can see that accuracy improved from 0.125 to 0.205 as informative genetic markers were included in the model. This improvement should be mainly attributed to the simulated association signals around the gene *MAP4* (the estimated effect size of rs11711953 is approximately −7.25 with the SE 1.46). For the real phenotype, although the performance was improved as a few genetic markers are included, the improvement was minor (R^2 increases approximately 1%). By checking the association signals, we were unable to detect significant associations.

For prediction using sequencing data, the model performed almost the same as that of GWAS data based on the simulated phenotypes. For the real phenotypes, the best accuracy achieved was close to R^2 = 0.285. Compared with the results of GWAS data, the prediction was not improved by using sequencing data.

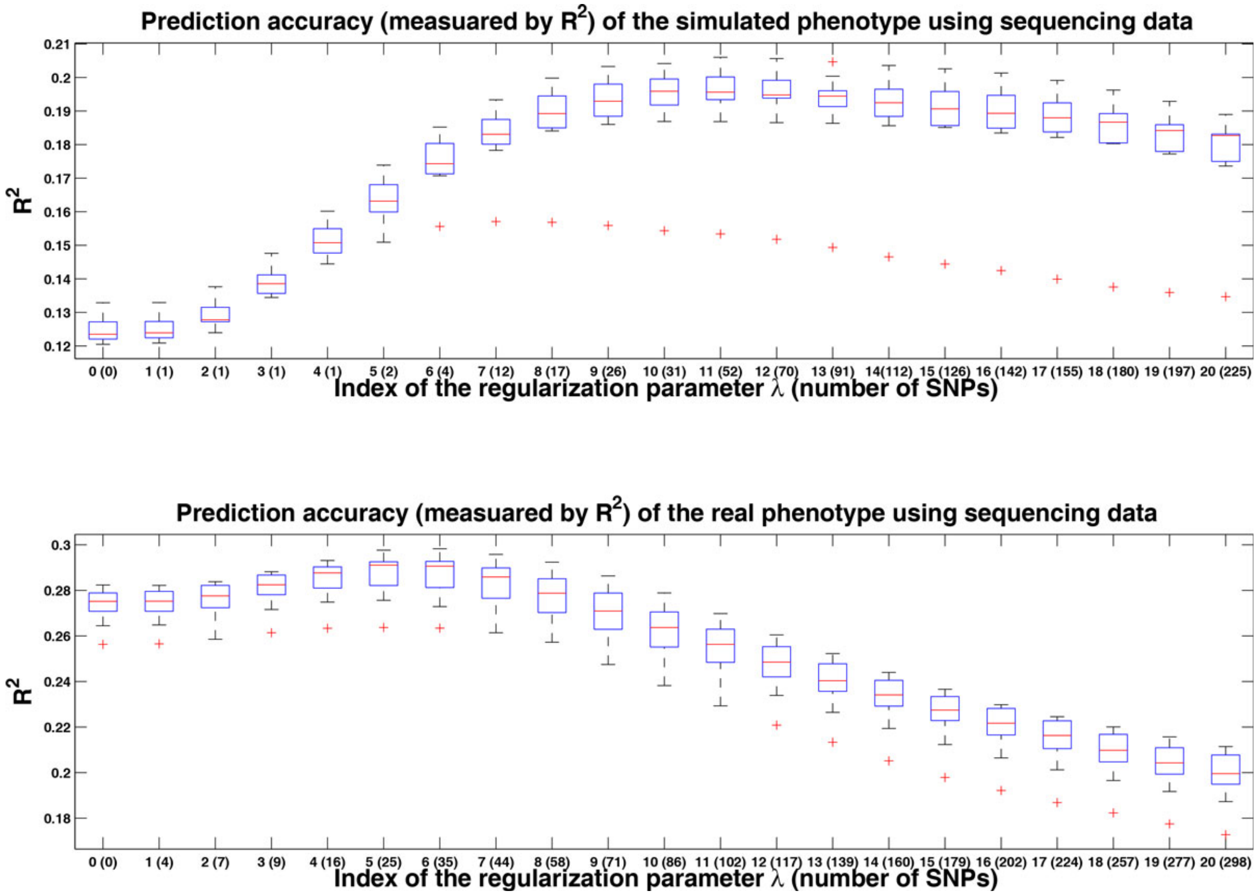

Figure 2 Prediction accuracy using sequencing data. Prediction results for the simulated phenotype and real phenotype using sequencing data. We ran 10-fold cross-validation 20 times to generate these boxplots. The x-axis is the index of the regularization parameter λ. The corresponding number of SNP markers used in the prediction model is given in brackets. Notice that these numbers are obtained using all samples. Index 0 corresponds to prediction only using pedigree information, thus the corresponding number of SNP marker is zero. The y-axis is the R^2 between the true value and the prediction value.

Discussion

In this study, we show that our model can integrate information from pedigree structures and genetic markers. This model will work well when there are a few markers with relatively large effect, as suggested by the analysis of the simulated phenotype. The reason is that the information from the pedigree structure is a global average of signals from the genetic background and the shared environmental influence. When there are some large effects that are different from the genetic background, it is better to extract them and consider them as fixed effects. In this way, the markers with larger effects can be treated locally. If all markers have similar effect sizes, the proposed method can only have minor improvement, as suggested by the analysis of the real phenotype.

In this study, we also compared the prediction performance based on GWAS data (ie, common variants) and sequencing data (ie, common variants plus low-frequency variants). The experimental result showed that inclusion of low-frequency variants could not lead to improvement of prediction accuracy. To significantly improve the prediction accuracy, the difficulty caused by polygenicity of complex traits needs to be addressed; that is, many small effects should be estimated more accurately. This implies that a larger sample size is needed. Besides the recruitment of more samples, an economic way is to combine multiple GWAS data sets of correlated traits. The underlying assumption is that these correlated traits may share some genetic factors. By borrowing information from each other, it is expected that the prediction accuracy can be dramatically improved.

Regarding the computational time, based on our current MATLAB implementation, we can finish a 10-fold cross-validation for the sequencing data in two hours.

Conclusions

In this study, genetic prediction can be improved by combining pedigree structure and information from genetic markers. We use a penalized LMM for this purpose and we show that it is computationally feasible. The experiment result based on the GAW18 data suggests that the main prediction power comes from the pedigree information. The additional improvement could be substantial if the effect sizes of a few genetic markers are noticeable, otherwise, could be minor. Integration of multiple GWAS data sets for genomic prediction may be a promising direction.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

CY and HZ designed the overall study; CY, CL, MC, XC and LH conducted statistical analyses; and CY and HZ drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank the organizers of GAW18 for providing the data set. We also thank Yale University Biomedical High Performance Computing Centre for computing resources, and NIH grant RR19895, which funded the instrumentation. CY, LH and HZ were supported in part by the NIH grants R01 GM59507, R01 DA12849 and R01 DA030976. CL, MC, and XC acknowledge support from CSC Scholarship.

The GAW18 whole genome sequence data were provided by the T2D-GENES Consortium, which is supported by NIH grants U01 DK085524, U01 DK085584, U01 DK085501, U01 DK085526, and U01 DK085545. The other genetic and phenotypic data for GAW18 were provided by the San Antonio Family Heart Study and San Antonio Family Diabetes/Gallbladder Study, which are supported by NIH grants P01 HL045222, R01 DK047482, and R01 DK053889. The Genetic Analysis Workshop is supported by NIH grant R01 GM031575.

This article has been published as part of *BMC Proceedings* Volume 8 Supplement 1, 2014: Genetic Analysis Workshop 18. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcproc/supplements/8/S1. Publication charges for this supplement were funded by the Texas Biomedical Research Institute.

Authors' details

[1]Department of Biostatistics, Yale School of Public Health, New Haven, CT 06520, USA. [2]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA.

Published: 17 June 2014

References

1. De Los Campos G, Allison DB: **Predicting genetic predisposition in humans: the promise of whole-genome markers.** *Nat Rev Genet* 2010, **11**:880-886.
2. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, *et al*: **Common SNPs explain a large proportion the heritability for human height.** *Nat Genet* 2010, **42**:565-569.
3. Visscher P, Yang J, Goddard ME: **A commentary on "Common SNPs Explain a Large Proportion the Heritability for Human Height" by Yang et al. 2010.** *Twin Res Hum Genet* 2010, **13**:517-524.
4. Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, Allison DB, de los Campos G: **Beyond missing heritability: prediction of complex traits.** *PLoS Genet* 2011, **7**:e1002051.
5. Gail MH: **Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk.** *J Natl Cancer Inst* 2008, **100**:1037-1041.
6. De Los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes J: **Predicting quantitative traits with regression models for dense molecular markers and pedigree.** *Genetics* 2009, **182**:375-385.
7. Gianola D, De Los Campos G, Hill WG, Manfredi E, Fernando R: **Additive genetic variability and the Bayesian alphabet.** *Genetics* 2009, **183**:347-363.
8. Schelldorfer J, Bühlmann P, van de Geer S: **Estimation for high-dimensional linear mixed-effects models using L_1-penalization.** *Scand Stat Theory Appl* 2011, **38**:197-214.
9. Rakitsch B, Lippert C, Stegle O, Borgwardt K: **A lasso multi-marker mixed model for association mapping with population structure correction.** *Bioinformatics* 2013, **29**:206-214.
10. Lippert C, Listgarten J, Liu Y, Kadie C, Davidson R, Heckerman D: **FaST linear mixed models for genome-wide association studies.** *Nat Methods* 2011, **8**:833-835.
11. Zhou X, Stephens M: **Genome-wide efficient mixed-model analysis for association studies.** *Nat Genet* 2012, **44**:821-824.
12. Tibshirani R: **Regression shrinkage and selection via the lasso.** *J R Stat Soc Series B Stat Methodol* 1996, **58**:267-288.
13. Zou H, Hastie T: **Regularization and variable selection via the elastic net.** *J R Stat Soc Series B Stat Methodol* 2005, **67**:301-320.
14. Mazumder R, Friedman J, Hastie T: **SparseNet: Coordinate descent with nonconvex penalties.** *J Am Stat Assoc* 2011, **106**:1125-1138.
15. Friedman J, Hastie T, Tibshirani R: **Regularized paths for generalized linear models via coordinate descent.** *J Stat Softw* 2010, **33**:1-22.

doi:10.1186/1753-6561-8-S1-S67
Cite this article as: Yang *et al.*: **A penalized linear mixed model for genomic prediction using pedigree structures.** *BMC Proceedings* 2014 **8** (Suppl 1):S67.